Sample solutions to Homework 2, Information-Theoretic Modeling (Fall 2014)

Jussi Määttä

September 18, 2014

Question 1

If you built your codebook by hand or took it from the web in the previous exercise, you now faced the challenge of writing your own code to build a codebook. As there are 27 allowed input characters (a-z and ' '), the number of two-character symbols is $27^2 = 729$.

Moreover, you should consider the case where your input has an odd number of characters. (The cleaned-up version of Chapter I of *Alice* has an even number of characters, so if you only used that input you might've missed this issue.) One solution is to append an end-of-file (EOF) character to odd-sized inputs. The sample code uses the character '#'. As a file may end with 27 different symbols, the total number of possible symbols becomes $27^2 + 27 = 756$.

The attached sample code is based on the same Huffman tree construction algorithm that we used earlier. Only minor modifications have been made to account for two-character symbols and the issue of odd-length input files. A shell script *test.sh* has been added for easy testing.

Like before, we "cheat" by constructing the codebook using our input. Now that we are using two-character symbols, not all symbols occur in the input (e.g. 'xz'). The program *build_codebook.py* now contains a Boolean flag *build_complete_codebook*, which determines whether we construct a codeword for each possible symbol or for only those that occur in our input.

How does this new approach affect the compression rate? Recall that last week, we managed to compress our input down to 43802 bits. Now, using two-character symbols, the input compresses down to 39064 bits (if *build_ complete_ codebook = False*) or 39065 bits (if *build_ complete_ codebook = True*). The input compresses to fewer bits than before.

Why did we get a better compression rate than last week? Because the individual letters in *Alice in Wonderland* are not really independent! For example, if we see the letter 'a', then the next letter is more likely to be 'l' than 'x'.

For completeness, let us roughly look at the sizes of the codebook files (*codebook.json*). They are stored quite inefficiently, of course, but let's just look at their sizes relative to each other. Last week's codebook (for single-character symbols) takes 403 bytes. The incomplete two-character codebook takes 6490 bytes and the complete codebook takes 106767 bytes. There is a price to pay for getting a smaller compressed file! We have hundreds of symbols, and because there are so many of them, we must have longer codelengths for each of them as well.

Question 2

(a)

First, recall that we can obtain the marginal distributions as follows:

$$p(X = x) = \sum_{y \in \{0,1,2\}} p(X = x, Y = y),$$
$$p(Y = y) = \sum_{x \in \{0,1\}} p(X = x, Y = y).$$

In particular, $p(X = 0) = 1/5 + 1/5 + 1/5 = 3/5$, $p(X = 1) = 0 + 1/5 + 1/5 = 2/5$, $p(Y = 0) = 1/5 + 0 = 1/5$, $p(Y = 1) = 1/5 + 1/5 = 2/5$ and $p(Y = 2) = 1/5 + 1/5 = 2/5$.

Now, by definition

$$\begin{aligned} H(X) &= -\sum_{x\in\{0,1\}} p(X=x)\,\log_2 p(X=x) \\ &= -p(X=0)\,\log_2 p(X=0) - p(X=1)\,\log_2 p(X=1) \\ &= -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5} \\ &\approx 0.9710 \end{aligned}$$

and

$$\begin{aligned} H(Y) &= -\sum_{y\in\{0,1,2\}} p(Y=y)\,\log_2 p(Y=y) \\ &= -\frac{1}{5}\log_2\frac{1}{5} - \frac{2}{5}\log_2\frac{2}{5} - \frac{2}{5}\log_2\frac{2}{5} \\ &\approx 1.5219. \end{aligned}$$

(b)

Here we also need the conditional probabilities $p(X = x \mid Y = y)$ and $p(Y = y \mid X = x)$. These are easy to read from the table or can be computed using the identity

$$p(X=x \mid Y=y) = \frac{p(X=x, Y=y)}{p(Y=y)}, \quad \text{when } p(Y=y) > 0.$$

We have

$$\begin{aligned} H(X \mid Y) &= -\sum_{y\in\{0,1,2\}}\sum_{x\in\{0,1\}} p(X=x, Y=y)\,\log_2 p(X=x \mid Y=y) \\ &= -p(X=0, Y=0)\,\log_2 p(X=0 \mid Y=0) \\ &\quad - p(X=0, Y=1)\,\log_2 p(X=0 \mid Y=1) \\ &\quad - p(X=0, Y=2)\,\log_2 p(X=0 \mid Y=2) \\ &\quad - p(X=1, Y=0)\,\log_2 p(X=1 \mid Y=0) \\ &\quad - p(X=1, Y=1)\,\log_2 p(X=1 \mid Y=1) \\ &\quad - p(X=1, Y=2)\,\log_2 p(X=1 \mid Y=2) \\ &= -\frac{1}{5}\log_2 1 - \frac{1}{5}\log_2\frac{1}{2} - \frac{1}{5}\log_2\frac{1}{2} - 0\log_2 0 - \frac{1}{5}\log_2\frac{1}{2} - \frac{1}{5}\log_2\frac{1}{2} \\ &= 0.8. \end{aligned}$$

(Side note: It may seem confusing that the above calculation included the term $p(X = 1, Y = 0) \log p(X = 1 \mid Y = 0)$ even though the "number" $0 \log 0$ is not really defined. However, this is a small technicality and we may just write $0 \log 0 = 0$. Why? We may resort to a limit argument, $\lim_{x \to 0+} x \log x = 0$. Or, strictly speaking, the double sum should be over the *support* of the random variable (X, Y), that is, the set $\{(x, y) \in \{0, 1\} \times \{0, 1, 2\} \colon p(X = x, Y = y) > 0\}$.)

Similarly,

$$\begin{aligned}
H(Y \mid X) &= -\sum_{x \in \{0,1\}} \sum_{y \in \{0,1,2\}} p(X = x, Y = y) \log_2 p(Y = y \mid X = x) \\
&= -p(X = 0, Y = 0) \log_2 p(Y = 0 \mid X = 0) \\
&\quad - p(X = 0, Y = 1) \log_2 p(Y = 1 \mid X = 0) \\
&\quad - p(X = 0, Y = 2) \log_2 p(Y = 2 \mid X = 0) \\
&\quad - p(X = 1, Y = 0) \log_2 p(Y = 0 \mid X = 1) \\
&\quad - p(X = 1, Y = 1) \log_2 p(Y = 1 \mid X = 1) \\
&\quad - p(X = 1, Y = 2) \log_2 p(Y = 2 \mid X = 1) \\
&= -\frac{1}{5} \log_2 \frac{1}{3} - \frac{1}{5} \log_2 \frac{1}{3} - \frac{1}{5} \log_2 \frac{1}{3} - 0 \log_2 0 - \frac{1}{5} \log_2 \frac{1}{2} - \frac{1}{5} \log_2 \frac{1}{2} \\
&\approx 1.3510.
\end{aligned}$$

Note that $H(X \mid Y) \neq H(Y \mid X)$.

(c)

Here we can simply use the identity $H(X, Y) = H(Y) + H(X \mid Y) \approx 1.5219 + 0.8 \approx 2.32$. Or equivalently, $H(X, Y) = H(X) + H(Y \mid X) \approx 0.9710 + 1.3510 \approx 2.32$.

(d)

$H(Y) - H(Y \mid X) \approx 1.5219 - 1.3510 \approx 0.17$.

(e)

We know that mutual information is symmetric: $I(X; Y) = I(Y; X)$. Now, by definition, $I(Y; X) = H(Y) - H(Y \mid X) \approx 0.17$.

Question 3

(a)

First, denote $\mathcal{X} = \{x_1, \ldots, x_r\}$, $\mathcal{Y} = \{y_1, \ldots, y_s\}$ and $\mathcal{Z} = \{x+y \colon x \in \mathcal{X}, y \in \mathcal{Y}\}$. By definition,

$$\begin{aligned} H(Z \mid X) &= -\sum_{x \in \mathcal{X}} \sum_{z \in \mathcal{Z}} p(X = x, Z = z) \log_2 p(Z = z \mid X = x) \\ &= -\sum_{i=1}^{r} \sum_{z \in \mathcal{Z}} p(X = x_i, Z = z) \log_2 p(Z = z \mid X = x_i). \end{aligned}$$

In the terms $p(X = x_i, Z = z) \log p(Z = z \mid X = x_i)$, we always have some fixed value x_i for the random variable X. Therefore, the random variable Z must take one of the values $x_i + y_j$, $j = 1, \ldots, s$. Hence, we may continue the above by writing

$$\begin{aligned} H(Z \mid X) &= -\sum_{i=1}^{r} \sum_{j=1}^{s} p(X = x_i, Z = x_i + y_j) \log_2 p(Z = x_i + y_j \mid X = x_i) \\ &= -\sum_{i=1}^{r} \sum_{j=1}^{s} p(X = x_i, X + Y = x_i + y_j) \log_2 p(X + Y = x_i + y_j \mid X = x_i). \end{aligned}$$

Now, notice that

1. $p(X = x_i, X + Y = x_i + y_j) = p(X = x_i, Y = y_j)$, since for the two events to occur simultaneously, we must have $X = x_i$ and thus the event $X + Y = x_i + y_j$ simplifies to $Y = y_j$.

2. $\log_2 p(X + Y = x_i + y_j \mid X = x_i) = \log_2 p(Y = y_j \mid X = x_i)$. Why? In general, for two events A and B with $p(B) > 0$, one can easily show that $p(A, B \mid B) = p(A \mid B)$. Then we can apply similar reasoning as above.

Thus, we have

$$\begin{aligned} H(Z \mid X) &= -\sum_{i=1}^{r}\sum_{j=1}^{s} p(X = x_i, Y = y_j)\, \log_2 p(Y = y_j \mid X = x_i) \\ &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(X = x, Y = y)\, \log_2 p(Y = y \mid X = x) \\ &= H(Y \mid X). \end{aligned}$$

Using the above result, we also have the following:

$$\begin{aligned} & X \text{ and } Y \text{ are independent} \\ \iff & I(Y; X) = 0 \\ \iff & H(Y) - H(Y \mid X) = 0 \\ \iff & H(Y) = H(Z \mid X) \\ \implies & H(Y) \leq H(Z), \end{aligned}$$

and $H(X) \leq H(Z)$ follows by symmetry. Here we used the fact that $H(A \mid B) \leq H(A)$ for all random variables A and B.

(b)

Let $p(X = 0) = p(X = 1) = 1/2$, that is, the random variable X represents a throw of an unbiased coin. Let $Y = -X$. Clearly X and Y are dependent: if we know the outcome of X, we can deduce the outcome of Y, and vice versa. A simple calculation shows that $H(X) = H(Y) = 1$. On the other hand, $Z = X + Y = X + (-X) = 0$, that is, Z is a constant: $p(Z = 0) = 1$. Hence, Z offers us "zero surprise", $H(Z) = 0$.

(c)

Assume that X and Y are independent. Furthermore, assume that given the value of Z, we can recover the values of X and Y. In other words, the mapping $f : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$, $f(x, y) = x + y$, is a bijection. We will show that these two assumptions imply $H(Z) = H(X) + H(Y)$, and that if either of them is violated, then equality is not guaranteed.

(Note: As discussed during and after the exercise session, if X and Y take a finite number of values, then the bijectivity of f is equivalent to $|\mathcal{X}| + |\mathcal{Y}| = |\mathcal{Z}|$.)

First, we prove a helpful inequality. Let A be a random variable and let g be a function so that $g(A)$ is also a random variable. Then, by the nonnegativity of entropy and the definition of joint entropy, we have

$$\begin{aligned} H(g(A)) &\leq H(g(A)) + H(A \mid g(A)) \\ &= H(A, g(A)) \\ &= H(A) + H(g(A) \mid A) \\ &= H(A) \end{aligned}$$

since $H(g(A) \mid A) = 0$ (if we know A, we can compute $g(A)$ so there is no surprise; $g(A)$ can only take one value once we know the value of A).

Apply then the above inequality to (X, Y):

$$\begin{aligned} H(Z) &= H(f(X, Y)) \\ &\leq H(X, Y) \\ &= H(X) + X(Y \mid X) \\ &= H(X) + H(Y) \end{aligned}$$

where we used the independence of X and Y.

Now, since we assumed that f is bijective, there exist functions g and h such that $X = g(Z)$ and $Y = h(Z)$. We can again apply the above inequality to the function $Z \mapsto (g(Z), h(Z))$ to obtain

$$\begin{aligned} H(X, Y) &= H(g(Z), h(Z)) \\ &\leq H(Z). \end{aligned}$$

We have shown that under our assumptions, $H(Z) \leq H(X) + H(Y) \leq H(Z)$, which means that $H(Z) = H(X) + H(Y)$.

To show that the independence of X and Y is required, let X be the flip of an unbiased coin and let $Y = X$. Given Z, we can always find out the values of X and Y, but X and Y are clearly dependent. We have $H(X) = H(Y) = 1$ and $H(Z) = H(X + Y) = H(2X) = H(X)$, so $H(Z) < H(X) + H(Y)$.

Finally, consider two random variables X and Y that represent independent flips of unbiased coins (with outcomes 0 or 1). Then $Z = X + Y$ may take one of three values: 0, 1 or 2. The probabilities are $p(Z = 0) = 1/4$, $p(Z = 1) = 1/2$, and $p(Z = 2) = 1/4$. If $Z = 1$, then we don't know whether

it was X or Y that came up heads. Our bijectivity assumption does not hold. Now, we already know that $H(X) = H(Y) = 1$, and

$$H(Z) = -\frac{1}{4}\log_2\frac{1}{4} - \frac{1}{2}\log_2\frac{1}{2} - \frac{1}{4}\log_2\frac{1}{4} = \frac{3}{2}$$

so $H(Z) \neq H(X) + H(Y)$.

Question 4

(a)

First, $H(X) = H(X_i) = -0.1\log_2 0.1 - 0.9\log_2 0.9 \approx 0.4690$.

Consider a sequence $\bar{x} = (x_1, \ldots, x_{100})$. Let k be the number of 1's in the sequence, that is, $k = \sum_{i=1}^{100} x_i$. We have $p(\bar{x}) = 0.1^k\, 0.9^{100-k}$ and therefore $\log_2 p(\bar{x}) = k\log_2 0.1 + (100-k)\log_2 0.9$.

Let $n = 100$ and $\varepsilon = 0.1$. The sequence $\bar{x}$ is in the typical set $A_\varepsilon^n = A_{0.1}^{100}$ if and only if

$$\begin{aligned} & 2^{-n(H(X)+\varepsilon)} \leq p(\bar{x}) \leq 2^{-n(H(X)-\varepsilon)} \\ \iff & -n(H(X)+\varepsilon) \leq \log_2 p(\bar{x}) \leq -n(H(X)-\varepsilon). \end{aligned}$$

We can solve these inequalities for k:

$$\begin{aligned} -n(H(X)+\varepsilon) &\leq \log_2 p(\bar{x}) \\ -100(-0.1\log_2 0.1 - 0.9\log_2 0.9 + 0.1) &\leq k\log_2 0.1 + (100-k)\log_2 0.9 \\ 10(\log_2 0.1 - \log_2 0.9 - 1) &\leq k\underbrace{(\log_2 0.1 - \log_2 0.9)}_{<0} \\ \underbrace{\frac{10(\log_2 0.1 - \log_2 0.9 - 1)}{\log_2 0.1 - \log_2 0.9}}_{\approx 13.2} &\geq k \end{aligned}$$

so we must have $k \leq 13$, and

$$\begin{aligned}
\log_2 p(\bar{x}) &\leq -n(H(X) - \varepsilon) \\
k \log_2 0.1 + (100 - k) \log_2 0.9 &\leq -100(-0.1 \log_2 0.1 - 0.9 \log_2 0.9 - 0.1) \\
k \underbrace{(\log_2 0.1 - \log_2 0.9)}_{<0} &\leq 10(\log_2 0.1 - \log_2 0.9 + 1) \\
k &\geq \underbrace{\frac{10(\log_2 0.1 - \log_2 0.9 + 1)}{\log_2 0.1 - \log_2 0.9}}_{\approx 6.8}
\end{aligned}$$

so we require $k \geq 7$.

Hence, the sequences in the typical set $A_{0.1}^{100}$ are those that have $7 \leq k \leq 13$ 1's.

(b)

The probability of the typical set $A_{0.1}^{100}$ is

$$p(A_{0.1}^{100}) = \sum_{k=7}^{13} \binom{100}{k} 0.1^k \, 0.9^{100-k} \approx 0.7590.$$

(At the exercise session, we discussed also the relative size of A_ε^n compared to the number of all possible lottery tickets 2^n. As it was pointed out, the ratio $|A_\varepsilon^n|/2^n$ becomes smaller and smaller as n increases. This can be seen by using the inequality $|A_\varepsilon^n| \leq 2^{n(H(X)+\varepsilon)}$. We have

$$\frac{|A_\varepsilon^n|}{2^n} \leq 2^{n(H(X)+\varepsilon)-n} = 2^{n(H(X)-1+\varepsilon)}.$$

Since coin flips are biased, we have $H(X) < 1$. Choose any $\varepsilon < 1 - H(X)$. Then $H(X) - 1 + \varepsilon < 0$ so the ratio tends to zero as $n \to \infty$.)

(c)

Here, we used $\sum_{k=7}^{13} \binom{100}{k} \approx 8.32 \cdot 10^{15}$ tickets to achieve a winning probability of about 0.76. If we had simply bought all tickets with at most twelve 1's, we would have obtain the winning probability

$$\sum_{k=0}^{12} \binom{100}{k} 0.1^k \, 0.9^{100-k} \approx 0.80$$

using $\sum_{k=0}^{12} \binom{100}{k} \approx 1.21 \cdot 10^{15}$ tickets—a better winning probability with fewer tickets.

The point here is that the strategy of buying the typical set becomes closer and closer to the optimum as n increases. For large enough n, the set $A_{0.1}^n$ has a probability of at least $1 - 0.1 = 0.9$, and the number of sequences in $A_{0.1}^n$ is about $2^{nH(X)}$. For large n, the probability becomes concentrated on the typical set and the "fringes" become less significant. (Consider, for example, the best ticket, i.e. the one with all zeros. Its probability of winning, 0.9^n, gets smaller as n increases.) Remember the informal version of the AEP from the lecture notes: "almost all sequences are almost equally likely".

Question 5

We know that for any two discrete random variables X and Y, we have $I(X;Y) \geq 0$ and equality holds if and only if X and Y are independent.

Let X be the roll of an eight-sided die where each of the outcomes $1, 2, \ldots, 8$ has the same probability. Then

$$H(X) = -\sum_{k=1}^{8} \frac{1}{8} \log_2 \frac{1}{8} = -\log_2 \frac{1}{8} = 3.$$

Similarly, let Y represent the roll of a 16-sided die; then $H(Y) = -\log_2(1/16) = 4$. The random variables X and Y have the desired entropies, and their independence implies $I(X;Y) = 0$.

On the other hand, we must have $I(X;Y) \leq \min\{H(X), H(Y)\}$. Why? Assume without loss of generality that $H(X) \leq H(Y)$. Then $I(X;Y) = H(X) - H(X \mid Y) \leq H(X)$ since $H(X \mid Y) \geq 0$.

Let X be the same random variable as before (a roll of an eight-sided die). Define the random variable $Y = (-1)^Z X$ where Z is a random bit with $P(Z = 0) = P(Z = 1) = 0.5$. Now, Y has 16 different outcomes, each with probability $1/16$, so $H(Y) = 4$. Moreover, $X = |Y|$, so $H(X \mid Y) = H(|Y| \mid Y) = 0$. Hence, $I(X;Y) = H(X) - H(X \mid Y) = H(X) = 3$.